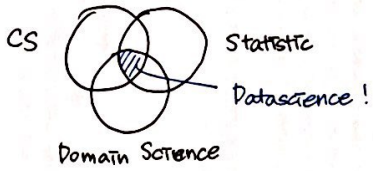
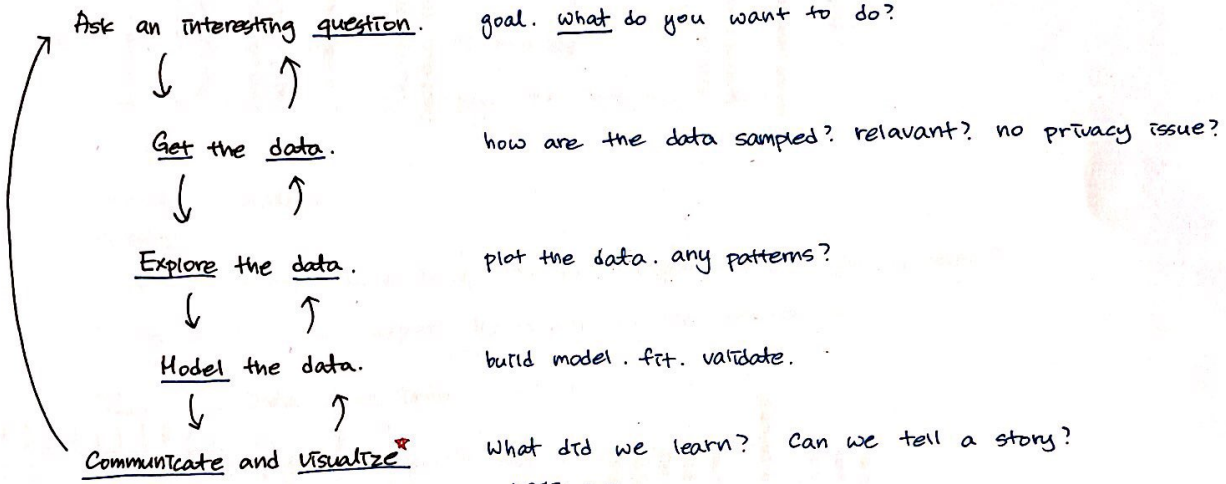
Lecture 1. What is data science?

- Data science :
  - Exploratory Data Analysis & Visualization
  - ML & Statistics
  - High-Performance Computing technologies for dealing with scale.



- Datascience Pipeline



- Computer Science vs. Real Science

- algorithm driven (focus on methods)	- data driven (focus on results, findings)
- try to build their own clean virtual world (try to invent things)	- try to understand messy natural world (try to discover things)
- random data can prove correctness of their methods	- "data" is important
- 8113 = 0.615 ——— : care about what the number is!	- 8113 ≈ 0.62 : care about what it means
- correct / wrong (둘 중 하나)	- Nothing is completely wrong.
- accuracy	- meaning

· Data Scientist

: real scientist 처럼 생각할 필요가 있다.

: are hired to produce insights ← come from understanding "meaning"

· Developing Curiosity

- Data scientists should develop curiosity about the domain / application they are working in.
  · broader perspective

· Asking Good Questions

- 주어진 data set 으로 부터 어떤 exciting 한 질문을 할 수 있는가?
- 무엇을 알고 싶어하는가? (사람들, 나 ..)
- 어떤 데이터셋이 필요한가?

· Practice Asking Questions : Who, What, Where, When, Why

1) Baseball-Reference.com : play 의 통계 정보.
   - Individual player's skill 을 측정하는 가장 좋은 방법은 무엇?
   - etc..

2) Demographic Questions
   (인구통계학의)
   - Do left-handed people have shorter lifespans than right handers?
   - Are heights and weights increasing in the population?

3) IMDb : Movie Data, Actor Data
   - Predict how many people will like a movie.
   - What is gross?
   - Social network of actors - (Six degrees of seperation)

4) NYC Taxi Cab data
   - Freedom of Information Act Request (FOIA) 에서 구할 수 있는 data.
   - How far do they travel?

5) Google Ngrams
   - Presents an annual time series of the frequency of every "popular" word/phrase with 1 to 5 words occurs in scanned books.
     ↑ appear more than 40 times
   - What is the lifespan of fame and technologies? Is it increasing or decreasing?

6) Google Trends
   - search term
   - example ?

· Properties of Data

1) Structured vs. Unstructured

Tweets from Twitter
Movie reviews (free text)
X-ray image
YouTube videos
→ unstructured data

Sales record spreadsheet
Financial Time series
→ structured data

2) Data Types

- Nominal (Categorical) (N)

: labels. = ≠

No order. 같다 / 다르다만 구분 가능. Index.

- Ordinal (O)

: ordered. = ≠ > <

- Quantatative (Q)

- Q-Interval

: location of zero arbitrary. = ≠ > < + −

only Intervals (distance) can be compared. Dates, Locations..

- Q-Ratio

: zero fixed. origin is important. = ≠ > < + − × ÷

can measure ratios & proportions. Length. Mass. Temperature..

· Classification vs. Regression

- Classification

: Assign a label ~~as~~ to an item from a discrete set of probabilities.

- Regression

: Forecast a numerical quantity (amount). the value is continuous.

Lecture 2. Mathematical Preliminaries

2.1 Probability

- Probability theory provides a formal framework for reasoning about the likelihood of events.
- **experiment** : a procedure which yields one ~~se~~ of a set of possible outcomes. rolling two dice
- **sample space S** : the set of possible outcomes of an experiment. $S = \{(1,1), (1,2), \cdots (6,6)\}$
- **event E** : a specified subset of the outcomes of an experiment. sum of dice = 7
  $E = \{(1,6), (2,5) \cdots (6,1)\}$
- probability of an outcome s : $p(s)$ where $\begin{cases} 0 \le p(s) \le 1 \\ \Sigma_{s \in S} p(s) = 1 \end{cases}$
- probability of an event E : $p(E) = \Sigma_{s \in E} p(s) = 1 - p(\bar{E})$
- **random variable V** : numerical function on the outcomes of a probability space.
  표본공간 S 를 정의역으로 하고, 실수를 공역으로 가지는 함수로 정의.
  확률변수를 사용하면 모든 표본을 실수에 대응시켜 표현할 수 있다.
  따라서 표본 공간 S 를 수직선 위에 표현할 수 있고 사건 E는 이 수직선 상 일부 구간으로 표현된다.
  $V((a,b)) = a+b$
  random variable V 를 sum of values of two dice 로 정의해볼 수 있다.
- **expected value** : 기대값. expected value of a random variable is defined as
  $E(V) = \Sigma_{s \in S} p(s) \cdot V(s)$

2.1.1 Probability vs. Statistics

- probability : predicting the likelihood of future event
- statistic : analyzes the frequency of past event
  applied to mathematics trying to make sense of observations (↳ past events) in the real world.

2.1.2 Compound Events and Independence

- probability of a compound event is combination of probability of two or more simple events.
- The events A and B are independent if and only if
  $P(A \cap B) = P(A) \times P(B)$
- Independence (zero correlation) : good to simplify calculations but
  bad for prediction ☹

2.1.3 Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$ B가 일어났을 때 A가 일어날 확률

- conditional pobabilities get interesting only when events are (NOT) independent.

∵ A, B가 independent 한 경우, $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$ ← 의미 없다!

- Bayes Theorem

: important tools which reverses the direction of the dependencies.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{P(A \cap B)}{P(B)} \cdot P(B)}{P(A)} = \frac{P(A|B)\,P(B)}{P(A)}$$

구하기 쉬운 것을 이용해서 구할 수 있다!

$$\Rightarrow \boxed{P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}}$$

구하기 어려울 때

ex) $P(spam|w) = \frac{P(w|spam) \cdot P(spam)}{P(w)}$

word 보고 spam 판단 → 어려움.

spam 보고 word 찾기 → 쉬움.

2.1.4 Probability Distributions

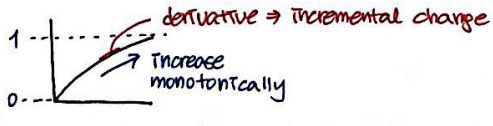
- Random variables are numerical functions where the values are associated with probabilities of occurrence.
- RVs can be represented by their probability density function (pdf).
- (pdf): x-axis : range of values the random variable can take on
  y-axis : probability of that given value.

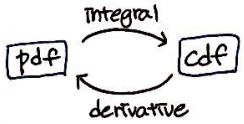
$$P(k = X) = \frac{h(k = X)}{\sum_x h(x = X)}$$

- pdf 는 histogram 과 유사하지만 y 축이 frequency 가 아니라 probability 값을 정의해 차이가 있다.
- RVs can also be represented by their cumulative density function (cdf)
- (cdf): running sum of the probabilities in the pdf. It reflects $P(X \le k)$ (↳ cdf) instead of $P(X = k)$ (↳ pdf)

→ may present misleading view of growth rate. (Apple 예시)

$$C(X \le k) = \sum_{x \le k} P(X = x)$$





: exact same information

2.2 Descriptive Statistics → ① representing data
② reducing data
→ data 에 대한 summary 제공, 대표값 값 제공.

- Descriptive statistics provides way to capture the properties of a given data set / sample

1) Central tendency measure : describe the center around which data is distributed

2) Variation or variability measures : describe data spread, i.e. how far the measurements lie from the center.

2.2.1 Centrality Measures

1) (Arithmatic) Mean

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

: Mean is meaningful for symmetric distributions without outliers, like height or weight.

(μ보다 작은 item 수) ≈ (μ보다 큰 item 수)

2) Median (중간값) : middle value.
Median is better for skewed distributions or data with outlier, like wealth

3) Mode (최빈값) : most frequent element in the data set.
Mode ~~is~~ often ~~is~~ is NOT close to the center.
text data 에서 자주 사용. 유사한 특징으로 grouping 할 때 mode 는 해당 cluster 를 대표하기 좋음.

4) Geometric Mean

$$\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}} = \sqrt[n]{a_1 a_2 \cdots a_n}$$ nth root of product of n values

: geometric mean <= arithmetic mean (always)

: geometric mean is sensitive to zero. (0 이나 0에 가까운 값의 영향이 큼!)

: Geometric means make sense with ratios. $\frac{1}{2}$, 2 의 geometric mean = 1

- Aggregation as Data Reduction

Representing a group of elements by a new derived element, like mean, min, count, sum reduces a large amount of dataset to a small summary statistic:
- reduce by a new derived element
- can become features when taken over natural groups or clusters in the full data set.

2.2.2 Variability Measures

- Standard Deviation

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}}$$ ← population 이때는 n, sample 이때는 n-1.

for large n, $n \approx (n-1)$ ∴ doesn't matter.

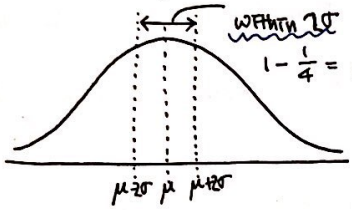
Variance = σ^2

- mean, standard deviation

together ⇒ characterize any distribution well

- parameterizing distributions

"Regardless of how data is distributed", at least $(1 - \frac{1}{k^2})$ th of the points must lie within k sigma of the mean.



⇒ tighter bounds apply for normal distribution.

normal 이때는 ±2σ 에 95%

2.2.3 Interpreting Variance

- Repeated observations of the same phenomenon do not always produce the same results, due to random noise or error.
- sampling error : when observations capture unrepresentative circumstances
  e.g. measuring rush hour traffic as on weekend as well as week days.
- measurement error : limits of precision inherent in any sensing device.
- signal to noise ratio : the degree to which a series of observations reflects a quantity of interest as opposed to data variance.
- hard to measure 'signal to noise ratio', because much of what you see is often variance due to sampling error and measurment error.

1) stock market : measuring the relative "skill" of different stock market investors.

2) sports performance : 선수의 실력이 시즌에 따라 변화함. → genuine difference..?

3) Model performance : simple model ~ complex model
↳ larger variance.
↳ more generalizable

2.3 Correlation Analysis

- Two factors are correlated when values of x has some predictive power on the value of y.
- correlation coefficient $r(X,Y)$: statistic that measures the degree to which Y is a function of X, and vice versa.
- correlation ranges from −1 to 1. −1 means anti-correlated
  0 means no relation
  1 means fully correlated

2.3.1 Correlation Coefficient : Pearson and Spearman Rank

There are two primary statistics used to measure correlation. These different statistics are appropriate in different situations.

1) The Pearson Correlation Coefficient ← more prominent

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

→ defines the degree to which a linear predictor of the form $y = m \cdot x + b$ can fit the observed data.
⇒ linear predictor 가 아닌 경우 의미없다.
e.g. $y = |x|$ ← correlation 0 나옴.

- X, Y are correlated (strongly)
  - X, Y 같은 방향 ⇒ positive ($X_i - \bar{X} > 0$ 일 때 $Y_i - \bar{Y} > 0$)
  - 반대 ⇒ negative
- X, Y are uncorrelated
  positive & negative terms should occur with equal frequency, offsetting each other.
  ⇒ value becomes zero.

numerator 에 있는. (determining sign of correlation) → named as covariance.

$$Cov(X,Y) = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})/n$$

denominator : reflects the amount of variance in the two variables, as measured by their standard deviation.

covariance between X and Y potentially increases with the variance of these variables

→ ~~따라~~ r 을 −1 ~ 1 로 맞추기 위함.

2) **Spearman Rank Correlation Coefficient** → better with non-linear relationships and outliers. less sensitive to outlier elements than Pearson.

- spearman rank correlation coefficient : counts the number of pairs of input points which are out of order.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$ → only cares about rank not the actual value

counts the number of disordered pairs, not how well the data fits. $-1 \le \rho \le 1$

$d_i = \text{rank}(x_i) - \text{rank}(y_i)$

$\text{rank}(x_i)$: the rank position of x_i in sorted order among all x_i.

$1 \le \text{rank}(x_i) \le n$

$p(x_1, y_{max}) \longrightarrow p(x_1, \infty)$: spearman 은 영향을 x (∵ rank(y) 변화하지 않아서!)
pearson ⇒ go crazy !

2.3.2 The Power and Significance of Correlation

- Correlation becomes more impressive the more points it is based on.

1) **Strength of correlation : R^2** ← square of correlation coefficient

r^2 estimates the fraction of the variance in Y explained by X in a simple linear regression.

→ Y 분산의 r^2 만큼만 X가 설명할 수 있다.

Let $f(x) = mx + c$ ← linear fit be predicted value of y from x.

Then, the residual value r_i is $r_i = y_i - f(x_i)$. (mean of $r = 0$)

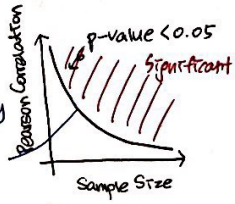
If X, Y are perfectly correlated, there will be no residual error. ∴ $V(r) = 0$

If X, Y are totally uncorrelated, the fit should contribute nothing ∴ $V(y) \approx V(r)$

for good linear fit $f(x)$, $V(r) < V(y)$

⇒ $1 - r^2 = V(r) / V(y)$

the level of correlation necessary to be statistically significance



Even small correlations become significant (p-value < 0.05) with large enough sample sizes.

2) **Statistical Significance**

The statistical significance of a correlation depends upon its "sample size n" as well as r. Traditionally, a correlation of n points is significant (if) there is an $\alpha \le \frac{1}{20} = 0.05$ chance that we would observe a correlation as strong as r in any random set of n points. * large # of weak but independent correlations may together have strong predictive power.

random 한 (무관한) r 만큼의 correlation을 관측할 확률이 0.05 미만일 때 n 개의 point 에서 관측한 correlation r 이 significant 하다. ↳ p-value

2.3.3 Correlation Does Not Imply Causation

- correlation does not mean causation. e.g. police, crime

2.3.4 Detecting Periodicities by Autocorrelation

- **Autocorrelation** : comparing a sequence to itself. → important concept in predicting future events.
- Autocorrelation function : series of correlations for all $1 \le k \le n-1$
- Time series data often exhibits cycles which affect its interpretation.
  A "cycle of length (k)" can be identified by unexpectedly large autocorrelation between S[t] and S[t+k] for all $0 < t < n-k$.
- Computing the lag-k autocorrelation takes O(n), but the full set can be computed in O(n log n) via the Fast Fourier Transform (FFT).

2.4 Logarithms

- logarithm : inverse exponential function.

$$y = \log_b x \implies b^y = x$$

① (Summing) logs of probabilities is more numerically stable than multiplying them.

$$\prod_{i=1}^{n} p_i = b^p \quad \text{where} \quad p = \sum_{i=1}^{n} \log_b(p_i)$$

ratio, power law 에 대해서 log 취하면 → good ~~distribution~~ analysis.

② Logarithms and (Ratios) ✓

- 비율이 2배 증가 하거나, $\frac{1}{2}$배로 감소하는 상황. ratio 는 4배 차이난다.
  log 를 취하면 +1.0, −1.0 으로 같은 크기의 차이로 표현할 수 있다.
- ratio 만 보면 outlier 처럼 보이는 것 → log 취해서 보면 outlier 가 아닐 수 있다!

③ Logarithms and (Power Laws) ✓

- power law distribution 에 log 취하면 traditional distribution 으로 보여줄 수 있다.

④ Normalizing (Skewed Distributions)